

PCTWORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68	A1	(11) International Publication Number: WO 99/43855 (43) International Publication Date: 2 September 1999 (02.09.99)
(21) International Application Number: PCT/US99/04094 (22) International Filing Date: 25 February 1999 (25.02.99) (30) Priority Data: 60/076,081 26 February 1998 (26.02.98) US (63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application US 60/076,081 (CIP) Filed on 26 February 1998 (26.02.98) (71) Applicant (for all designated States except US): GENOMICS COLLABORATIVE, INC. [US/US]; Building 700, Suite 2200, One Kendall Square, Cambridge, MA 02139 (US). (72) Inventors; and (75) Inventors/Applicants (for US only): BING, David, H. [US/US]; 27 Waverly Street, Brookline, MA 02146 (US). WILLIAMSON, Janice, M. [US/US]; 104 Green Street, Wakefield, MA 01880 (US). (74) Agents: BROOK, David, E. et al.; Hamilton, Brook, Smith & Reynolds, P.C., Two Militia Drive, Lexington, MA 02421 (US).		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG). Published <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>
(54) Title: UNIQUE IDENTIFIER FOR BIOLOGICAL SAMPLES		
(57) Abstract The present invention provides a method for internal labelling of a biological sample by which the sample is identifiably linked to its source and other relevant information, based on the polymorphisms inherent in the sample itself. A set of polymorphisms in the sample is detected, and the resulting data is used as a unique identifier which is then used to identify the sample. This unique identifier can also be used to identify the source of the sample, and any other relevant information.		

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

UNIQUE IDENTIFIER FOR BIOLOGICAL SAMPLES

RELATED APPLICATION(S)

This application claims priority to application 60/076,081, filed February 26, 1998, the entire teachings of which are incorporated herein by reference.

5 BACKGROUND OF THE INVENTION

With the advent of the Human Genome Project and the advances in technology that have resulted, biological and genetic testing have become increasingly more common. Hospitals and other health care entities are using new tests for diseases, and are processing more samples for testing than ever before. The ease and speed of many biological tests has also increased enormously, so that these tests are now being widely used outside of the health care industry. Veterinarians, of course, have always closely followed advances in human health care. But law enforcement agencies now routinely employ DNA-based methods in forensics, and even population geneticists, ecologists, and evolutionary biologists use these methods to track the evolution and variability within and between populations of organisms.

When handling large numbers of samples, accurate and reliable tracking of samples and quality control of associated information is vital. In hospital settings, aberrant test results are always a cause for concern because doubts are then cast on the state of the patient's health. In a tissue repository, it must be possible for a sample (or portions of a sample) to be reliably and repeatedly retrieved with no doubts as to the sample's identity. Mislabeling or loss of labeling of a sample may mean that the sample is rendered useless if it cannot be accurately connected back to

the sample's history and/or source. Most samples and their sources are given a common alphanumeric designation, and this designation is also linked to information about the source and the sample (e.g., patient name, sample type, disease condition, etc.). A loss of this designation from its association with either
5 the source, or the sample, or the information will often result in a complete loss of utility of all three.

This potential loss of association between the designation and the sample is especially likely in settings where very large numbers of samples are being processed. Machine errors, while problematic, generally result in the destruction of
10 large numbers of samples, and so are noticed easily. Human error, however, has the potential to cause serious errors that go unnoticed for a period of time. These include transcription errors, misplacing or swapping of samples, destruction of labels, off-by-one errors (resulting in a series of samples where the designation or information from each sample is misassociated with the next sample). In addition,
15 pages from lab notebooks can be obliterated or lost, and magnetic media corrupted. Databases containing all of this information can be backed up, but intervening data added to the database since the last backup is usually lost. If an error is introduced and not discovered until after a backup is made, then this error effectively replaces the "true" data. In addition, many facilities save only the most recent backup, or
20 store backups at the same site as the current data, resulting in loss of all information in the event of a physical disaster (e.g. a fire).

SUMMARY OF THE INVENTION

The present invention relates to a method of creating a unique identifier for reliably identifying samples, their sources, and associated information. The use of
25 the identification system described herein substantially decreases potential mixups and misidentification of samples, their sources, and associated information.

Specifically, the present invention provides a method for creating a unique identifier which is used to label the sample, its source, or the associated information, based on the polymorphisms inherent in the sample and its source. One or more
30 polymorphisms in the sample is detected, and the resulting polymorphism data is

used to produce a unique identifier, which is then used to identify the sample. This unique identifier can also be linked with the source, and/or any information that may be associated with either the sample or the source (*i.e.* the unique identifier can be used as a common designation for the sample, its source and/or other relevant information). If this unique identifier is separated from the sample, then the polymorphisms within the sample simply need to be re-detected to reproduce the polymorphism data which is then used to produce the unique identifier, thereby recreating the proper unique identifier, and, ultimately, its link to its source.

In general, the invention features a method for producing a unique identifier for a biological sample, comprising detecting one or more polymorphisms within the biological sample, and selecting one or more polymorphisms sufficient to form a unique identifier. The biological sample can be from a vertebrate, an invertebrate, a plant, or consist of microorganisms. The biological sample can also be from a mammal, particularly a human. The sample can be blood, saliva, hair, body fluid, tissues, organs, one or more cells, or a whole organism. The polymorphisms can be nucleic acid polymorphisms, protein polymorphisms, enzyme polymorphisms, chemical polymorphisms, biochemical polymorphisms, phenotypic polymorphisms, and quantitative polymorphisms, particularly a nucleic acid sequence polymorphism, a nucleic acid length polymorphism, or a short tandem repeat (STR). The unique identifier can also be linked to the source of the biological sample, or relevant information about the biological sample or the source of the biological sample. The unique identifier can be in the form of an alphanumeric string, or a bar code.

The invention also features a method for establishing a repository containing a collection of biological samples, comprising obtaining a biological sample from a source, detecting one or more polymorphisms in the sample, selecting one or more polymorphisms sufficient to form a unique identifier, using the unique identifier to identify the sample, storing the sample with the unique identifier, and repeating these steps for biological samples from other sources. The samples, in general, are DNA-containing samples, particularly from humans, and the polymorphisms are nucleic acid polymorphisms, protein polymorphisms, enzyme polymorphisms, chemical polymorphisms, biochemical polymorphisms, phenotypic polymorphisms,

and quantitative polymorphisms, or short tandem repeat (STR). The unique identifier can be in the form of an alphanumeric string, or a bar code, and can also be linked to the source of the biological sample, or relevant information about the biological sample or the source of the biological sample.

5 In addition, the invention features a method of determining, by means of a unique identifier, if a source is represented by a sample within the repository, comprising obtaining a sample from the source, detecting one or more polymorphisms in the sample selecting one or more polymorphisms sufficient to form a unique identifier, and comparing the unique identifier so produced to the
10 unique identifier of each sample in the repository, where shared identity between the two unique identifiers indicates that the source is already represented in the repository. In general, the samples are DNA-containing samples, preferably from humans. The polymorphisms are nucleic acid polymorphisms, particularly short tandem repeats (STR), protein polymorphisms, enzyme polymorphisms, chemical
15 polymorphisms, biochemical polymorphisms, phenotypic polymorphisms, and quantitative polymorphisms. The unique identifier can also be linked to the source of the biological sample, or relevant information about the biological sample or the source of the biological sample. The unique identifier can be in the form of an alphanumeric string, or a bar code.

20 The invention also features a method for linking, by means of a unique identifier, a first biological object lacking a unique identifier with a second object having a unique identifier, comprising detecting one or more polymorphisms in the first biological object, selecting one or more polymorphisms sufficient to form a unique identifier, and comparing the unique identifier so made to the unique
25 identifier of the second object, where shared identity between the two unique identifiers links the first biological object with the second object. The biological sample can be from a vertebrate, an invertebrate, a plant, or consist of microorganisms. The biological sample can also be from a mammal, particularly a human. The sample can be blood, saliva, hair, body fluid, tissues, organs, one or
30 more cells, or a whole organism. The polymorphisms can be nucleic acid polymorphisms, particularly short tandem repeats (STR), protein polymorphisms,

enzyme polymorphisms, chemical polymorphisms, biochemical polymorphisms, phenotypic polymorphisms, and quantitative polymorphisms, particularly a nucleic acid sequence polymorphism, or a nucleic acid length polymorphism. The unique identifier can also be linked to the source of the biological sample, or relevant
5 information about the biological sample or the source of the biological sample. The unique identifier can be in the form of an alphanumeric string, or a bar code.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method for creating a unique identifier for identifying biological samples, their sources, or associated information based on the
10 polymorphisms inherent in the sample and source themselves. In this method, the nucleic acid contained within the sample itself is used to produce the unique identifier for identifying and linking the sample, its source, and associated information. No external material needs to be added to the sample which could dilute or alter the accuracy of other test results. An advantage of the invention,
15 therefore, is that it is unnecessary to add "identifying sequences" to the samples, and that without such additions, one may conduct studies of genetics, disease associations, evolutionary relationships, etc., without the results being tainted by the added identifying sequences.

A "source" or "the source from which the sample is derived" refers to the
20 originating material for a sample. A source of a biological sample, for example, can be a human, any animal, plant, insect, or a population or strain of microorganisms. A source of a biological sample does not have to be living, and can be a deposit in a tissue repository, herbaria or museum specimens, forensic specimens, or fossils. A "potential source" as used herein, means a source from which the sample may
25 possibly have been taken in the past.

By "sample" is meant a portion of source biological material that originated elsewhere, *i.e.*, the sample was removed from its source. A sample can be any biological sample, (*e.g.*, blood, saliva, hair, organs, biopsies, bodily fluids, one or more cells), and can be taken from any vertebrate, including mammals such as
30 humans, or plant, insect, reptile. The sample can also be a strain or mixed

population of microbes. Samples can also be biological materials taken from defunct or extinct organisms, *e.g.*, samples can be taken from pressed plants in herbarium collections, or from pelts, taxidermy displays, fossils, or other materials in museum collections.

5 "Information associated with" the sample or source from which the sample is derived is meant to include, without limitation, any information that might be necessary or advantageous to be linked to the sample or the sample source, *e.g.* name, address, sex, medical history (in the case of human samples), species, collection data, provenance (in case of non-human samples), etc.

10 Once a biological sample is taken from its source, the sample is tested across one or more polymorphic loci, and the polymorphic data produced are used to create a unique identifier, which is identifiably linked to the sample, and serves as its unique designation. This unique identifier can also be identifiably linked to the sample's source, and/or any information that may exist concerning the source and/or
15 the sample. By saying that the unique identifier is "identifiably linked" to the sample, sample source, or related information means that it is connected in some way with any or all of these three things, *e.g.* the unique identifier may be on a label attached to a container holding the sample, the unique identifier may exist as a field in a database record containing medical data regarding the source, etc. In essence,
20 the genetic code of the sample itself, which is unique and forms the basis of the polymorphisms tested, serves as the unique identifier. Because the unique identifier is based on the genetic code, which is unique between individuals, the unique identifier will also be unique between samples from different source individuals.

A "polymorphism" is an allelic variation between two samples. As used
25 herein, the term includes differences between proteins (*e.g.*, enzymes, blood groups, blood proteins), differences in the chemicals and biochemicals (*e.g.*, secondary metabolites) produced by the source organism(s), differences between nucleic acids involving differences in the nucleotide sequence (*e.g.*, restriction site maps), or differences in length of a stretch of nucleic acid (*e.g.*, RFLPs (restriction fragment
30 length polymorphisms), microsatellites, STRs (short tandem repeats), SSRs (simple sequence repeats), SSLPs (simple sequence length polymorphisms), VNTRs

(variable number tandem repeats)). Allelic variation can also result in phenotypic (*i.e.*, visually- apparent) polymorphisms, or variations in quantitative characters (*e.g.* variation in height, length, yield of fruit, etc.) between the organisms that serve as the source of the samples. With some types of biological material, phenotypic differences may be visible in the samples themselves, *e.g.*, kernels of different types of "Indian" corn often appear very different from each other, with red, yellow, white, blue, streaked kernels, etc. With such samples, phenotypic polymorphisms could also be used to produce the unique identifier.

A polymorphism is not limited by the function or effect it may have on the organism as a whole, and can therefore include allelic differences which may also be a mutation, insertion, deletion, point mutation, or structural difference, as well as a strand break or chemical modification that results in an allelic variant. A polymorphism between two nucleic acids can occur naturally, or be caused intentionally by treatment (*e.g.*, with chemicals or enzymes), or can be caused by circumstances normally associated with damage to nucleic acids (*e.g.*, exposure to ultraviolet radiation, mutagens or carcinogens).

As used herein, a "sequence polymorphism" is a difference in the sequence of two nucleic acids or two amino acids. Two amino acid sequences can differ by having different residues at a particular position (*i.e.*, and amino acid substitution), or some residues may be deleted, or new residues inserted or added to one or more ends. Two nucleic acids differing in sequence may have the same number of base pairs (*e.g.*, "ATGC" vs. "ATTC"), but may also include some differences in overall sequence length as well (*e.g.*, "ATCACATG" vs. "ATCACACATG"). Types of commonly-studied polymorphisms caused by sequence differences include restriction site polymorphisms, isozymes, differences in protein conformation, and length polymorphisms. If the nucleic acid is sequenced, then a sequence difference itself (as represented by the string of letters) serves as the polymorphism.

As used herein, a "length polymorphism" is a difference in the length of two nucleic acids. Two different nucleic acids with a length polymorphism between them also have a sequence polymorphism, but many methods used to detect a length polymorphism do not reveal the exact sequence polymorphism. Commonly-used

types of length polymorphisms include RFLPs (restriction fragment length polymorphisms), microsatellites, STRs (short tandem repeats), SSRs (simple sequence repeats), SSLPs (simple sequence length polymorphisms), and VNTRs (variable number tandem repeats).

5 In general, the difference between "length polymorphisms" and "sequence polymorphisms" is generally in the methods used to detect them. With RFLPs, for example, restriction endonucleases are used to cut a nucleic acid molecule into fragments, which are then separated on an agarose gel. The differences between two individuals are measured by the changes in size of the resultant nucleic acid
10 fragments, and so are referred to as length polymorphisms, yet those differences are *caused* by differences in the underlying sequence, which is the basis for the change in restriction sites, and therefore the changes in the sizes of the nucleic acid fragments. Because the method of detection/visualization can only differentiate on the basis of fragment length, the RFLPs are generally classed as length
15 polymorphisms.

As used herein, a "polymorphic locus" is a segment of nucleic acid which may contain a polymorphism as described above. It is not required that the precise sequence of the nucleic acid be known. A polymorphic locus is not limited to those loci which are polymorphic in all situations, *e.g.*, a polymorphic locus which
20 displays an allelic variation between individuals A and B, but not between individuals A and C, remains a polymorphic locus for purposes of comparing individuals A and B, as well as individuals B and C.

"Nucleic acid" means deoxyribonucleic acid (DNA), ribonucleic acid (RNA), nucleic acids from mammals or other animals, plants, insects, bacteria,
25 viruses, or other organisms.

By "unique identifier" is meant an identification tag, designation, or code to be linked to a sample, its source, or other information, such as patient case history, disease testing results, genetic testing results, geographic or temporal collection data, or any other information which may be useful when linked with the sample or
30 source. The unique identifier can exist in the form of an alphanumeric string, a bar

code, an entry in a database, or any other useful human-readable or machine-readable form.

The extent to which such an identifier will be unique depends on the loci chosen for polymorphism testing. Allelic polymorphism has been studied for
5 decades, and there are many genetic systems which have been commonly used in assessing polymorphism between populations or individuals. These include classical blood groups, blood proteins, isozymes, distribution of restriction endonuclease sites, restriction fragment length polymorphisms (RFLPs), and others. The most successful to date, however, have been microsatellites, also known as short
10 tandem repeats (STRs), or simple sequence length polymorphisms (SSLPs), or variable number tandem repeats (VNTRs).

STRs are stretches of DNA that consist of repeated sequences repeats. The base sequence is usually just a few base pairs long, typically two to twelve base pairs, but longer base repeats have been seen. This base sequence is then tandemly
15 repeated, and the number of times it is repeated can vary greatly, depending on the STR locus being studied. An STR can therefore be expressed as $(X)_n$, where X is the repeated sequence, (e.g. "CA") and n is the number of times that X is repeated.

Most individuals in a population will have the same STR at the same location in the genome, that is different individuals will have the same base repeat
20 at the same location, but the precise number of repeats often varies from individual to individual. For example, for a given STR mapped to a particular location on the genome, the base sequence may be repeated 5 times in individual A, but may be repeated 8 times in individual B and 20 times in individual C.

These tandem repeats are believed to be caused by "slippage" of the DNA
25 polymerase enzyme as the DNA is replicated. In general, n increases over generations, and the amount of slippage varies over time and in different lines. The variability of these repeated sequences is generally correlated to the length of the base repeat, with STRs composed of longer base repeats exhibiting less variability between individuals than shorter base repeats. For example, a two base pair repeat
30 may consist of a two base pair unit being repeated hundreds of time in an individual, while a 12-base pair unit may only be repeated a few times. In general, the amount

of slippage that occurs during replication, and therefore the amount of variability in the number of repeats that results from that slippage, is also correlated to the length of the base repeat. Short repeats tend to exhibit higher rates of polymorphism between individuals, while 10- or 12-base pair repeats may show little or no
5 variability.

STRs can be amplified and detected by known procedures. For example, they can be detected by electrophoretic separation followed by radionuclide or fluorescent labeling, or silver staining. They have many advantages over other methods of detecting polymorphisms (*e.g.*, RFLPs) because of their small size, the
10 ease and speed with which they can be detected and analyzed, and the fact that the process is amenable to automation. The more recent generations of large-scale genetic maps have been made using STRs (Hudson, T.J. *et al.*, *Science* 270:1945-1954 (1995); Dietrich, W.F., *et al.*, *Nature Genetics* 7:220-245 (1994); Yerle, M., *et al.*, *Mamm. Genome* 6:176-186 (1995); Jacob H.J., *et al.*, *Nature Genetics* 9:63-69
15 (1995)). Because of the extremely high rate of polymorphism of some of the STR loci, they are also used in forensic tests by law enforcement agencies.

A number of kits for amplifying STR loci are commercially available, and the rates of polymorphism of these loci in different ethnic backgrounds are known. These include AmpFISTR™ Profiler, AmpFISTR™ Profiler Plus, AmpFISTR™
20 Green I (PE Applied Biosystems, Foster City, California, USA), the Geneprint™ STR Systems (Promega Corp., Madison, WI), including Geneprint™ PowerPlex™ 1.1, Geneprint™ PowerPlex™ 1.2, Geneprint™ PowerPlex™ 2, and Geneprint™ PowerPlex™ 16, Sex Determination Systems, and others. These STR systems were developed for use in humans, but microsatellite markers have been developed in
25 other organisms, including horse, cattle, sheep, goat, dog, pig, mouse, rat, barley, corn, soybean, and others.

These loci can be used singly, or can be combined, depending on the power of discrimination required. As the number of organisms being studied and the number of individuals from which samples are removed and archived increases, the
30 degree of polymorphism required to uniquely identify each sample also increases,

and the number of polymorphic loci that need to be tested to have a sufficient number to create the unique identifier also increases.

For example, if three individuals possessed the following alleles at three different loci:

5		Locus 1	Locus 2	Locus 3
	Individual A	1,3	1,2	3,5
	Individual B	2,3	2,2	1,4
	Individual C	2,3	1,5	3,5

then detection of the alleles at Locus 1 would allow a sample from Individual A to
10 be distinguished from a sample from B or C, but samples from B and C could not be uniquely distinguished from each other, and a second locus would need to be tested. On the other hand, Locus 2 alone could serve as the unique identifier, because by itself, it can serve to distinguish between samples from all three individuals.

The Power of Discrimination (P_D) of a given system of loci is defined as the
15 probability that two individuals selected at random will differ with respect to that given system of loci. The P_D is related to the Probability of Identity (P_I) by the equation

$$P_D = 1 - P_I,$$

Where P_I is determined by solving the equation

20
$$P_I = \sum X_i^2,$$

where X_i is the frequency in the population of the i th allele. The allelic frequencies within different ethnic populations are known for many of the polymorphisms of the STR loci in the commercially-available kits, so a set of STRs which will provide a unique identifier for every sample can be chosen, even if the final number of
25 samples is not known. Combinations of loci can be chosen that have matching probabilities of less than 1 in several million or more (See, for example Table 1).

Table 1. Matching probabilities of various populations in the Geneprint™ STR system, using fluorescent detection (Promega Corp., Madison, Wisconsin, USA).

	Caucasian-American	African-American	Hispanic-American
CTTv Quadriplex	1/6623	1/25575	1/7194
FFFL Quadriplex	1/2632	1/16807	1/3279
5 Both combined	1/17,400,000	1/430,000,000	1/23,600,000

Once the polymorphism rates are known for a series of loci, one can choose which loci and how many will go into making up the unique identifier. If it is anticipated that the final number of samples will be relatively small, than only a few loci are sufficient to form the unique identifier, and one or more of the loci may not
 10 need to have a high P_D . On the other hand, if one intends to store a very large number of samples, then it would be prudent to use more loci, each with a high P_D . The loci selected will be based on considerations such as P_D , anticipated size of the repository, ease of use, applicability to the organisms being sampled, cost, and availability.

15 Polymorphisms used

The polymorphisms that can be used in the invention will vary depending on the types of samples being stored. STRs are well-studied in humans, and kits are commercially available for amplifying a number of STR loci. Genetic maps based on STRs have been built for other organisms (*e.g.*, mouse, rat, pig). STRs appear to
 20 exist in most higher organisms, and are easy to isolate and characterize. Because the methods used to identify and assess STRs are virtually identical for different organisms, one skilled in the art can isolate STRs in an organism of choice, assess the polymorphism rates, and choose those most useful in the present invention. For many organisms, STRs and their primer sequences have been published in the
 25 scientific literature. One wishing to use previously published STRs need only order those primers (*e.g.*, custom primers can be ordered and received in 48 hours from Research Genetics, Huntsville, Alabama, USA), and then use them to amplify the STRs in the DNA of the collected samples.

It is not necessary to use commercially-available primers to practice the present invention, nor is it necessary to use microsatellite markers developed by others. The present invention allows one to use any polymorphic marker that is convenient, so long as it provides a power of discrimination between individuals.

- 5 There are many species worthy of study for which no genetic map exists. One of the reasons that microsatellite markers have become so successful is that they are easy to develop for previously unstudied organisms. One already familiar with an organism for which there are no microsatellite markers can develop them with relative ease using methods well-known in the art.

10 Uses of the invention

The method described herein will be of particular use in a pathology laboratory or testing facility, or a large-scale cryogenic repository. Maintaining the integrity of the sample labels is of paramount importance in these situations, as quality control problems often result from failure of the record-keeping system.

- 15 Naturally, such a method will also be of use to blood banks, tissue banks, and veterinary hospitals and testing facilities.

- The method can also be used by large repositories to identify misplaced or misidentified samples. For example, a tissue bank may take a small piece (*e.g.*, a sample) of a stored tissue (*e.g.*, a source) for testing (*e.g.*, tissue typing for a potential recipient of the tissue). If the identification were disassociated from the sample (*e.g.*, the label fell off the test tube), those test results would normally be lost. Using the unique identifier of the present invention, however, one would simply test the sample for the polymorphic loci, and recreate the unique identifier. The sample (and the tissue typing test results) could then be reassociated with the source in storage.
- 20
25

- The method is especially useful to maintain the long-term integrity of samples and associated information, especially in tissue repositories. Many biomedical studies require analysis of tissue samples from large populations of individuals with known medical, dietary, genetic, social, and cultural backgrounds.
- 30 During the course of a study requiring several years to complete, it may be necessary

to test a particular sample several times. It is therefore vital to the accuracy of the study to confidently re-retrieve that sample.

Many biomedical studies involve analysis of a group of individuals with a set of characteristics in common (*e.g.*, cigarette smoking, ethnic background, incidence of particular cancers). At present, the amount of time and effort involved in assembling a set of individuals appropriate for a study may be greater than the effort of conducting the study itself. If samples from a large number of individuals could be collected in a repository along with associated data on the individuals, then it should be possible to "assemble" a set of individuals for a given study by selecting samples from these individuals chosen on the basis of a defined set of characteristics. For example, if a blood sample repository contained samples from 100,000 individuals, and associated medical data on those same individuals in computerized form, then a medical study could be conducted by selecting individuals with desired characteristics (as listed in the computerized medical data), and then retrieving samples (or more likely, sub-samples) from those individuals, which are held in the repository. The method of the present invention is useful in establishing such a repository, because the method greatly reduces the likelihood of samples being misidentified and allows confident re-retrieval of samples.

Another advantage of the invention is that if the unique identifier is de-associated from the sample within the repository (*e.g.*, the label falls off the tube) analysis of the polymorphisms in the sample allows re-creation of the unique identifier.

Use of the method of the invention also provides a method for preventing repository deposit of samples from duplicate sources, because when the unique identifier is created for a sample from a new source, one need only search the repository records for that same unique identifier to see if the source is already represented by a sample in the repository.

The invention also has uses outside of the medical field. Because of the increasing ease with which samples from various sources (*e.g.*, plant, animal, microbial, fungal, viral) can be tested for polymorphisms, the invention is applicable in any situation where a large number of biological samples may be stored. An

example of such a situation would be a field study of biodiversity of a wild population. New tests for assessing diversity (*i.e.*, assessing polymorphism) are continually being created, and samples from previous collecting expeditions represent a "snapshot" of the biodiversity that existed in the past. Such previously-
5 collected samples can be re-tested using current techniques, but the results are only useful if the integrity of the sample designations is still sound, and the samples can be linked to their original collection data. Maintenance of quality of the record keeping is especially important if the field samples are from species which are endangered or extinct. The method of the invention provided here has potential uses
10 in studies of population genetics, evolutionary genetics, and ecology. In studies of flora and fauna from locales that are either increasing or decreasing in pollution, for example, it is necessary to both store the samples for a period of time and also maintain their identification. Such sampling at periodic intervals is also a requirement of an effective bioremediation plan.

15 There are many situations where one would want to keep a biological sample for a period of time against the possibility of testing it again later. For example, even if one has conducted a population genetics study on a series of samples (collection of organisms), a new test developed at a future time may allow the testing of different hypotheses, and provide the answer to new questions, without
20 necessitating collection of new samples in the field. Therefore, the method of the invention described herein would be especially useful in maintaining collections of samples from endangered species. The unique identifier and identity of each sample can be re-verified from the sample itself.

This sample identification method can be used to keep track of samples in
25 any study or collection where there are a large number of biological samples being stored for a period of time, and where there is a chance that samples may become misplaced or mislabeled.

EXAMPLES

Example 1: Use of STRs to Create a Unique Identifier

A biological sample is obtained from a human, and an aliquot is taken for polymorphism testing. DNA is isolated by methods well known in the art (Maniatis *et al.*, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York; Ausubel, F.M. *et al.*, eds., *Current Protocols in Molecular Biology*). An amount of this isolated DNA is removed, GenePrint™ primers (Promega Corp., Madison, WI) for the CSF1PO locus are added to it, and amplification is carried out, all according to the manufacturer's instructions (supplemental information on thermocycling are well known in the art, see e.g., Innis, M.S., *et al.* (1990) *PCR Protocols: A Guide to Methods and Applications*, Academic Press, Inc. San Diego, CA). After amplification, fluorescence detection is carried out, also according to the manufacturer's recommendations. The process is repeated for the other loci in the CTTv Quadriplex (TPOX, TH01, vWA), and also the four loci in the FFFL Quadriplex (F13A01, FESFPS, F13B, and LPL).

Once detection of the polymorphism(s) is complete, the unique identifier can be created for the sample. For a system such as these eight loci, where the alleles are 3 to 7 repeats in length, a convenient conversion method is to simply list each locus by letter, followed by the two alleles for that locus. For a sample with alleles of 3 and 5 tandem repeats at the first locus, alleles of 2 and 7 repeats at the second, etc., the unique identifier would be "A35B27...HXY". The precise conversion method could be varied, depending on the number of repeats in the loci, e.g., a locus with 3 - 12 repeats would require 4 digits after the locus letter.

Example 2: Preparation of DNA from Samples of Whole Blood

Red blood cells lack DNA because they are enucleated, and must therefore be lysed to facilitate their separation from white blood cells, which contain genomic DNA. After the red blood cells are lysed and removed, the white blood cells are then lysed with an anionic detergent in the presence of a DNA stabilizer, which limits the activity of DNase. Contaminating RNA is then degraded with RNase, and the RNA, proteins, and other contaminants are then removed by salt precipitation. The genomic DNA is recovered by alcohol precipitation, dissolved in TE buffer, and stored. Because the genomic DNA will be used in a nucleic acid amplification

method, it is advisable to also have a "blank" control tube (containing reagents but no blood) accompany the blood sample tube through the extraction process. After extraction, the "DNA" from the "blank" control tube would be amplified to ensure that no extraneous DNA has contaminated the extraction process

- 5 Isolation of genomic DNA from whole blood can be accomplished by following any of a variety of protocols, including using the PUREGENE[®] kit (Gentra Systems, Minneapolis, Minnesota, USA), and following the manufacturer's instructions. Place 30 ml of RBC Lysis Solution into a 50 ml tube, add 7 ml to 10 ml of whole blood, mix by inverting several times, and incubate for 10 minutes at
- 10 room temperature. Invert the tube again at least once during the incubation. Centrifuge the tube for 10 minutes at 2,000 x g, pour off the supernatant, leaving behind the visible white cell pellet and about 200 µl of residual liquid. Vortex the tube vigorously for 20 seconds to resuspend the cells in the residual liquid. Add 10 ml of Cell Lysis + RNase A (made fresh that day), and vortex on high speed for 10
- 15 seconds. Incubate the tube at 37°C for 15 to 30 minutes to allow digestion of the RNA.

- Cool the sample to room temperature by placing in an ice bath for 10 minutes. Add 3.33 ml of the Protein Precipitation Solution (Gentra Systems, Minneapolis, Minnesota, USA) into the tube. Vortex at high speed for 20 seconds to
- 20 mix uniformly. Centrifuge at 2,000 x g for 10 minutes. If a tight, dark brown pellet is not formed, repeat the 20-second vortex, followed by a 5-minute incubation on ice, and repeat the 10-minute centrifugation at 2,000 x g.

- Pour off the supernatant into a clean 50 ml tube containing 10 ml of 100% isopropanol. Mix by inverting gently 50 times (do not vortex, or the DNA will be
- 25 sheared). The DNA is stable at this point, and can be stored indefinitely in the isopropanol.

- Centrifuge at 2,000 x g for 3 minutes. Carefully pour off the supernatant, leaving behind the white pellet, and drain the tube upside down on clean absorbent paper. Add 10 ml of 70% ethanol, and wash the pellet by inverting gently, avoiding
- 30 dislodging the pellet. Centrifuge at 2,000 x g for 1 minute. Carefully pour off the

ethanol, leaving the pellet behind. Invert carefully so as to not dislodge the pellet, and drain the tube on clean absorbent paper for 10 minutes.

- Add 1 ml of DNA Hydration Solution (Gentra Systems, Minneapolis, Minnesota, USA), and rehydrate the DNA by incubating at 65°C for 1 hour, 5 overnight at room temperature, and at 65°C 1 hour the next day. Tap the tube periodically to help disperse the DNA. The DNA in solution can be stored indefinitely at 4°C.

Example 3: Preparation of Blood Sample With FTA™ Paper

- A blood sample is drawn from a human. Two µl of blood are placed on a 10 piece of FTA™ paper (FITZCO, Inc., Maple Plain, Minnesota, USA), dried, and stored until ready to be processed.

- To analyze the polymorphisms in the sample, a 1 mm disc is punched directly into a 2 ml microcentrifuge tube, and 200 µl of FTA™ purification reagent is placed on the disc. The tube is capped, vortexed for 3-5 seconds, then centrifuged 15 in a microcentrifuge at 12,000 x g for 30 seconds. The wash solution is then aspirated and discarded. The wash is then repeated with another 200 µl of purification reagent. After the second wash solution has been aspirated and discarded, the disc is washed twice with TE as follows: 200 µl of TE buffer is added, and the disc vortexed for 3-5 seconds, the tube and disc are then centrifuged 20 at 12,000 x g for 30 seconds and the filtrate removed and discarded. After the disc has been washed twice with TE, the disc is subjected to polymorphism analysis.

Example 4: Analysis of Polymorphisms in a Blood Sample

- A 1 mm punch of FTA™ paper containing a blood sample, processed as described in Example 2, *supra*, is placed in a 0.5 ml tube, and tested with the 25 AmpFISTR Profiler Plus™ system (Perkin Elmer Applied Biosystems, Foster City, California, USA), according to the manufacturer's instructions. In general, to the tube is added 10.5 µl of Profiler Plus Reaction Mixture, 0.5 µl of Taq Gold, and 5.5 µl of Primer Mixture. The tube is sealed, and placed in a thermocycler under the following conditions: 95°C for 11 minutes, followed by 24 cycles of: 94°C for 1

minute, 59°C for 1 minute, 72°C for one minute. After the 25th cycle, the reaction mixture is placed at 60°C for up to 83 minutes. After thermocycling is complete, the reaction is held at 4°C until ready for gel electrophoresis.

Example 5: Analysis of Amplification Products by Gel Electrophoresis

- 5 Five µl of amplification product (produced as described above in Example 3) are mixed with 5 µl of 2X loading buffer (0.25% Bromphenol Blue, 12.5% Ficoll 400, 50 mM EDTA, 5X TAN (10X TAN: 0.4 M Tris, 40 mM Na Acetate Trihydrate, 10 mM EDTA, pH to 7.9 with acetic acid)). The 10 µl mixture is loaded into a well in a 1% agarose gel prepared with TE buffer and containing 0.5 µg of
- 10 ethidium bromide per ml of agarose gel. Appropriate size ladder is also loaded on the gel. The gel is then electrophoresed in TAE buffer for 1 hour at 100 volts, and then illuminated with UV light on a transilluminator, and photographed. The bands in the photograph are then compared to the literature supplied by the manufacturer to determine the precise alleles present in the sample.

15 Example 6: Creation of the Unique identifier

A set of blood samples was prepared and tested with the AmpF/STR Profiler Plus™ system as described in Examples 2 through 4, and the results are shown in Table 2.

20 Table 2. Alleles found in seven human individuals when tested for eight STR loci in the AmpF/STR Profiler Plus™ system (PE Applied Biosystems, Foster City, California, USA).

Locus	#1	#2	#3	#4	#5	#6	#7
D3S1358	15	14,15	15,18	16,17	15,17	17,18	15
vWA	15,18	15,16	13,14	16,17	15,20	17,19	14,15
25 FGA	19,24	20,21	22,24	23	21,22	21,23	24,28
Amelogenin	X,Y	X,Y	X,Y	X,Y	X	X	X
D8S1179	12,15	13,15	12,14	13,15	14	13	14,15
D21S11	33.2	29,30	30,31.2	29,32	28,31	28,33.2	32.2,38

D5S818	12	11,13	8,12	1012	13,14	11	12,13
D13S317	12,13	11,13	11,12	9,12	12,14	9,14	12
D7S820	8,9	8,12	10,11	8,11	10,12	10,11	9,10
D18S51	11,15	9,14,2	13,2,20	18,24	9,18	10,12	10,18

- 5 The polymorphism data for a sample can be coded in a number of ways. The raw data for individual #1, for example, is as follows:

D3S1358 15;vWA 15,18;FGA 19,24;Amelogenin X,Y;D8S1179 12,15;
D21S 11,33.2;D5S818 12;D13S317 12,13;D7S820 8,9;D18S51 11,15

- This data can be used "raw" as the unique identifier (*i.e.*, "as is," as above), with no
10 alteration. For repositories with very large numbers of samples, this may be desirable, as it is the most "foolproof" method.

Alternatively, the STR loci can be "coded," *i.e.*, each locus represented by a combination of numbers or letters, *e.g.*, D3S1358 can be represented by "A" or "01," vWA by "B" or "02," etc. The raw data so coded would then be:

- 15 A,15,B,15,18,C,19,24,D,X,Y,E,12,15,F,33.2,G,12,H,12,13,I,8,9,J,11,15, or
01,15,02,15,18,03,19,24,04,X,Y,05,12,15,06,33.2,07,12,08,12,13,09,8,9,10,11,15

- All patents, patent applications, and references cited above are hereby incorporated by reference in their entirety. While this invention has been particularly shown and described with references to preferred embodiments thereof,
20 it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention as defined by the appended claims.

CLAIMS

What is claimed is:

1. A method for producing a unique identifier for a biological sample, the method comprising:
 - 5 (a) detecting one or more polymorphisms within the biological sample; and
 - (b) selecting one or more polymorphisms sufficient to form a unique identifier;thereby producing a unique identifier for a biological sample.
- 10 2. The method of Claim 1, wherein the biological sample is taken from an organism selected from the group consisting of: vertebrates, invertebrates, plants, and microorganisms.
3. The method of Claim 1, wherein the biological sample is from a mammal.
4. The method of Claim 3, wherein the mammal is a human.
- 15 5. The method of Claim 3, wherein the biological sample is selected from a group consisting of blood, saliva, hair, body fluid, tissues, organs, and one or more cells.
6. The method of Claim 5, wherein the polymorphism is selected from the group consisting of nucleic acid polymorphisms, protein polymorphisms,
20 enzyme polymorphisms, chemical polymorphisms, biochemical polymorphisms, phenotypic polymorphisms, and quantitative polymorphisms.

7. The method of Claim 6, wherein the polymorphism is a nucleic acid sequence polymorphism.
8. The method of Claim 7, wherein the polymorphism is a nucleic acid length polymorphism.
- 5 9. The method of Claim 8, wherein the polymorphism is a short tandem repeat (STR).
- 10 11. The method of Claim 1, wherein the unique identifier is also linked to relevant information about the biological sample or the source of the biological sample.
12. The method of Claim 1, wherein the unique identifier is selected from the group consisting of an alphanumeric string, and a bar code.
- 15 13. A method for establishing a repository containing a collection of biological samples, wherein each biological sample has a unique identifier associated with it, the method comprising:
 - (a) obtaining a biological sample from a source;
 - (b) detecting one or more polymorphisms in the sample;
 - (c) selecting one or more polymorphisms sufficient to form a unique
20 identifier;
 - (d) using the unique identifier to identify the sample;
 - (e) storing the sample with the unique identifier;
 - (f) repeating steps (a) through (e) for biological samples from other sources;

-23-

thereby establishing a repository containing a collection of biological samples, wherein each such biological sample has a unique identifier associated with it.

14. The method of Claim 13, wherein the samples are DNA-containing samples.
- 5 15. The method of Claim 13, wherein the sample source is a human.
16. The method of Claim 13, wherein the polymorphism is selected from the group consisting of nucleic acid polymorphisms, protein polymorphisms, enzyme polymorphisms, chemical polymorphisms, biochemical polymorphisms, phenotypic polymorphisms, and quantitative
10 polymorphisms.
17. The method of Claim 16, wherein the polymorphism is a short tandem repeat (STR).
18. The method of Claim 13, wherein the unique identifier is selected from the group consisting of an alphanumeric string, and a bar code.
- 15 19. The method of Claim 13, wherein the unique identifier is also linked to the source of the biological sample.
20. The method of Claim 13, wherein the unique identifier is also linked to relevant information about the biological sample or the source of the biological sample.
- 20 21. A method of determining, by means of a unique identifier, if a source is represented by a sample within the repository of Claim 15, the method comprising:
 - (a) obtaining a sample from the source;

- (b) detecting one or more polymorphisms in the sample;
 - (c) selecting one or more polymorphisms sufficient to form a unique identifier, wherein the polymorphisms selected are those used to form unique identifiers for the samples within the repository;
 - 5 (d) comparing the unique identifier of (c) to the unique identifier of each sample in the repository;
- wherein shared identity between the unique identifier of (c) to a unique identifier of a sample in the repository indicates that the source is represented by a sample within the repository.
- 10 22. The method of Claim 21, wherein the samples are DNA-containing samples.
23. The method of Claim 21, wherein the source of the sample is a human.
24. The method of Claim 21, wherein the polymorphism is selected from the group consisting of nucleic acid polymorphisms, protein polymorphisms, enzyme polymorphisms, chemical polymorphisms, biochemical
- 15 polymorphisms, phenotypic polymorphisms, and quantitative polymorphisms.
25. The method of Claim 24, wherein the polymorphism is a short tandem repeat (STR).
26. The method of Claim 21, wherein the unique identifier is also linked to the
- 20 source of the biological sample.
27. The method of Claim 21, wherein the unique identifier is also linked to relevant information about the biological sample or the source of the biological sample.

28. The method of Claim 21, wherein the unique identifier is selected from the group consisting of an alphanumeric string, and a bar code.
29. A method for linking, by means of a unique identifier, a member of a first group with a member of a second group, wherein the first group comprises a biological sample lacking a unique identifier, and a source of a biological sample lacking a unique identifier, and wherein the second group comprises a biological sample having a unique identifier, a source of a biological sample having a unique identifier, or information having a unique identifier, the method comprising:
- 10 (a) detecting one or more polymorphisms in the member of the first group;
 - (b) selecting one or more polymorphisms sufficient to form a unique identifier, wherein the polymorphisms selected are those used to form the unique identifier for the members of the second group;
 - 15 (c) comparing the unique identifier of the member of the first group to the unique identifier of the member of the second group;
- wherein shared identity between the unique identifier of the member of the first group and the unique identifier of member of the second group, links the member of the first group with the member of the second group.

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/US 99/04094

A. CLASSIFICATION OF SUBJECT MATTER

IPC 6 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No
X	US 5 556 955 A (VERGNAUD GILLES) 17 September 1996 see the whole document ---	1-29
X	WO 92 13969 A (BAYLOR COLLEGE MEDICINE) 20 August 1992 see the whole document ---	1-29
Y	WO 96 19586 A (DUNN JAMES M ;VISIBLE GENETICS INC (CA)) 27 June 1996 see the whole document ---	1-29
Y	WO 96 10648 A (PROMEGA CORP) 11 April 1996 see the whole document ---	1-29
P.Y	WO 98 14275 A (INTRONN LLC ;MITCHELL LLOYD G (US)) 9 April 1998 see the whole document ---	1-29
-/--		

☒ Further documents are listed in the continuation of box C

☒ Patent family members are listed in annex

Special categories of cited documents

A document defining the general state of the art which is not considered to be of particular relevance

E earlier document but published on or after the international filing date

L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

O document referring to an oral disclosure, use, exhibition or other means

P document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance, the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

8 document member of the same patent family

Date of the actual completion of the international search

28 June 1999

Date of mailing of the international search report

05/07/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5318 Patentlaan 2
NL - 2230 HV Rijswijk
Tel: (+31-70) 340-2040, Tx: 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Molina Galan, E

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/US 99/04094

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>SHUBER A P ET AL: "HIGH THROUGHPUT PARALLEL ANALYSIS OF HUNDREDS OF PATIENT SAMPLES FOR MORE THAN 100 MUTATIONS IN MULTIPLE DISEASE GENES" HUMAN MOLECULAR GENETICS, vol. 6, no. 3, 1 March 1997, pages 337-347, XP000673446</p> <p>-----</p>	

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 99/04094

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5556955 A	17-09-1996	FR 2680520 A	26-02-1993
WO 9213969 A	20-08-1992	US 5364759 A	15-11-1994
		AU 670231 B	11-07-1996
		AU 1334392 A	07-09-1992
		CA 2101859 A	01-08-1992
		EP 0639228 A	22-02-1995
WO 9619586 A	27-06-1996	US 5776737 A	07-07-1998
		AU 4113696 A	10-07-1996
		CA 2208428 A	27-06-1996
		DE 19581886 T	07-05-1998
		GB 2311370 A,B	24-09-1997
WO 9610648 A	11-04-1996	AU 3998195 A	26-04-1996
		CA 2118048 A	31-03-1996
		US 5843660 A	01-12-1998
WO 9814275 A	09-04-1998	AU 4738697 A	24-04-1998